BASE: a Bias-Aware news Search Engine for improving user awareness [Prototype]

Monica Lestari Paramita[1,*], Maria Kasinidou[2], Styliani Kleanthous[2] and Frank Hopfgartner[1,3]

[1]*University of Sheffield, Sheffield, United Kingdom*
[2]*Open University of Cyprus, Nicosia, Cyprus*
[3]*Universität Koblenz, Koblenz, Germany*

Abstract

The BASE prototype aims to improve user awareness of biases in search engine results. It utilises existing resources and NLP tools to identify biases in news articles. It incorporates bias visualisation features to inform users of biases in each news article and at the search results level. It also incorporates results reranking features to allow users to retrieve different sets of results based on their search preferences. Preliminary evaluation results suggest the prototype achieves a positive usability score (64.3 out of 100) and has a potential for increasing user awareness of biases, with the reranking features rated more useful than the bias visualisation features.

Keywords

bias in search engines, interface design, evaluation

1. Introduction

Increasingly, it becomes obvious that news search engines may include biases in their search results [1]. These biases may appear at the *article level*, e.g., an article may present a view that is politically biased to a certain political ideology (e.g., left wing). In other cases, an article may produce a certain *focus*, e.g., a report on COVID-19 rate for a specific country, or an article on COVID-19 vaccine for a specific manufacturer. The focus of the article may not necessarily introduce bias in the content itself, e.g., an article that focuses on Pfizer does not necessarily presents a view that is *biased* towards Pfizer. However, if a query 'covid vaccine' retrieves mostly articles with Pfizer as the entity focus, this may be seen as a bias at the *results level*. Biases at the results level may also be caused by search engine's localisation, which promotes search results with the same geographical focus as the users' location [2]. Although localisation aims to provide relevant results, these results also highly limit users' views of the topic, often without users' awareness of the results that they do *not* see. The lack of user awareness of these

DESIRES 2022 – 3rd International Conference on Design of Experimental Search & Information REtrieval Systems, 30-31 August 2022, San Jose, CA, USA

[*]Corresponding author.
m.paramita@sheffield.ac.uk (M. L. Paramita); maria.kasinidou@ouc.ac.cy (M. Kasinidou); styliani.kleanthous@gmail.com (S. Kleanthous); hopfgartner@uni-koblenz.de (F. Hopfgartner)
0000-0002-9414-1853 (M. L. Paramita)

biases have been shown to manipulate users' understandings of a topic [3] and influence their decision making [4].

Previous studies have proposed a number of visualisations to increase user awareness of biases. News aggregators, such as AllSides [5] and GroundNews [6], have presented news articles that represent multiple political ideologies to provide users with a balanced view. Hamborg et al. [7] provides matrix-based results to support users in accessing news events from news publishers in different locations (as they often present different perspectives). Other studies, such as Papadakos and Konstantakis [8], have also explored the importance of displaying biased aspects for the entire search results. However, very few studies have investigated designs that visualises *multiple* types of biases, which are often the case for news articles.

In this paper, we introduce a novel prototype of a search engine interface designed to increase users' awareness of multiple types of biases in the results. The prototype also aims to provide the ability to users to easily access different facets of the results. Instead of developing new methods for measuring biases, the prototype makes use of available resources and techniques to inform users of possible biases in the results. This means that such system can be made usable in the near future to support users in their information seeking tasks. An initial evaluation of how users respond to these visualisations are also provided in this study. This work provides a valuable contribution in understanding how bias-aware news search engines should be designed.

2. BASE: Bias-Aware news Search Engine prototype

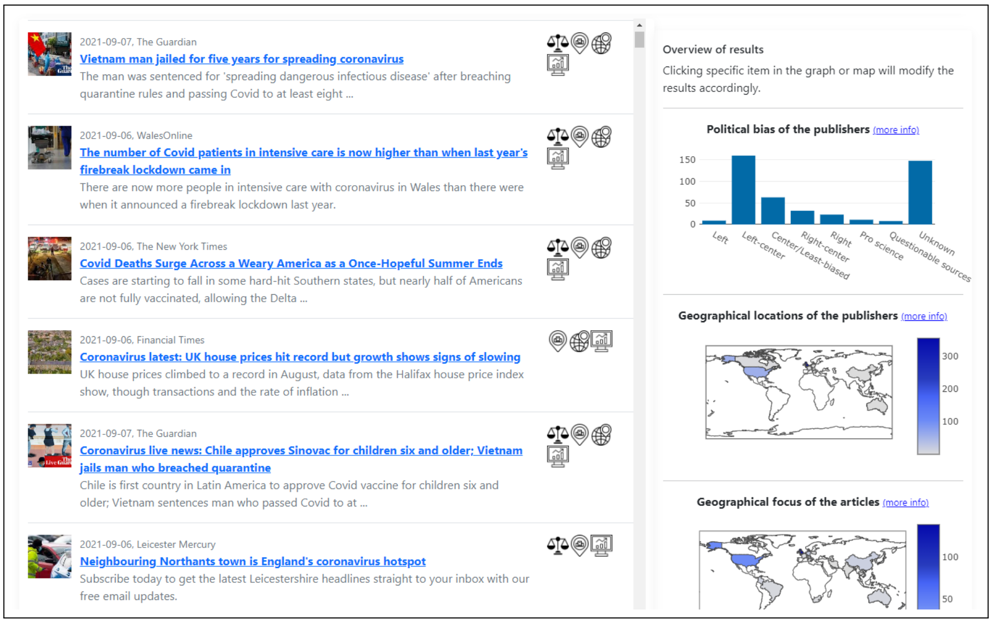
2.1. Design

To identify specific features to include in the design, we conducted three user studies on designing bias-aware search engines using a participatory approach. These resulted in eight designs that incorporated two different approaches: i) *bias visualisation approach*, for informing users of possible biases in the results, and ii) *results-reranking approach*, which allows users to access different results by modifying (the ranking of) the results. We invited 132 participants to evaluate these eight designs. The findings suggest that users would like i) to see information on different types of biases in search results, ii) the ability to retrieve a different set of results using their preferred aspect, and iii) to have both approaches in search engines.

We incorporated findings from these studies into the design of *BASE*.[1] The prototype provides both bias visualisation and results-reranking features. As proof-of-concept, we selected four aspects to be included in the re-ranking features: political bias, geographical locations of the publishers, geographical focus of the articles, and the entity focus of the articles. More aspects may be integrated in the next future if methods to measure them become available.

When users access the prototype, they are asked to enter a query (e.g., "coronavirus") to start searching. Once the query is submitted, the system will display the search results (Figure 1), showing a list of relevant articles in the left panel. In addition to the articles, the prototype shows two *bias visualisation features*. The first feature provides bias information at the *article level* (shown in the left panel as different icons on the right side of each article). Each icon represents different types of biases. When a user hovers on an icon (e.g., the scale), it provides

[1] https://cycat.group.shef.ac.uk/prototype/BASE/



2021-09-07, The Guardian
Vietnam man jailed for five years for spreading coronavirus
The man was sentenced for 'spreading dangerous infectious disease' after breaching quarantine rules and passing Covid to at least eight ...

2021-09-06, WalesOnline
The number of Covid patients in intensive care is now higher than when last year's firebreak lockdown came in
There are now more people in intensive care with coronavirus in Wales than there were when it announced a firebreak lockdown last year.

2021-09-06, The New York Times
Covid Deaths Surge Across a Weary America as a Once-Hopeful Summer Ends
Cases are starting to fall in some hard-hit Southern states, but nearly half of Americans are not fully vaccinated, allowing the Delta ...

2021-09-06, Financial Times
Coronavirus latest: UK house prices hit record but growth shows signs of slowing
UK house prices climbed to a record in August, data from the Halifax house price index show, though transactions and the rate of inflation ...

2021-09-07, The Guardian
Coronavirus live news: Chile approves Sinovac for children six and older; Vietnam jails man who breached quarantine
Chile is first country in Latin America to approve Covid vaccine for children six and older; Vietnam sentences man who passed Covid to at ...

2021-09-06, Leicester Mercury
Neighbouring Northants town is England's coronavirus hotspot
Subscribe today to get the latest Leicestershire headlines straight to your inbox with our free email updates.

Overview of results

Clicking specific item in the graph or map will modify the results accordingly.

Political bias of the publishers (more info)
150
100
50
0
Left
Left-center
Center/Least-biased
Right-center
Right
Pro science
Questionable sources
Unknown
Geographical locations of the publishers (more info)
300
200
100
Geographical focus of the articles (more info)
100
50

Figure 1: BASE prototype interface

information on the type of biases and the specific biased aspect of the article (e.g., “Political bias: left-center”). The second feature provides bias information at the *results level* (shown in the right panel) in the form of bar charts and choropleth maps. These visualisations show the distribution of political biases of the publishers, geographical location of the publishers, geographical focus of the articles and the entity focus of articles in the search results (see Figure 3).

These visualisations also incorporated *results-reranking features*. By clicking a specific aspect in the figures, users can easily obtain a new set of results containing articles only from the specified political bias, country, or entity. E.g., clicking “Left-center” on the political bias bar chart will retrieve only articles from news publishers identified to have a “left-center” bias. Similarly, by clicking on “Australia” in the “geographical focus of the articles” map, users will be able to view only those articles reporting COVID-19 in Australia. We describe the methods to identify and visualise these biases in Section 2.2.

2.2. Workflow of the BASE prototype

This section describes the information processing workflow of the search engine (illustrated in Figure 2) and outlines the methods used to measure and visualise these biases.

We limited our index on news articles related to the COVID-19 pandemic. For this, we used the most popular queries for this topic according to Google Trends in February 2021. We retrieved 100 news articles per query returned by Google News using the Zenserp API [9]. This process was conducted daily to allow users to access the most updated news articles.

For each article, we carried out two processes. Firstly, we extracted the URL of the publishers for the news articles, e.g. `bbc.co.uk` (BBC), or `ft.com` (Financial Times). These URLs were then used to identify the political bias of the publishers and the location of the publishers.

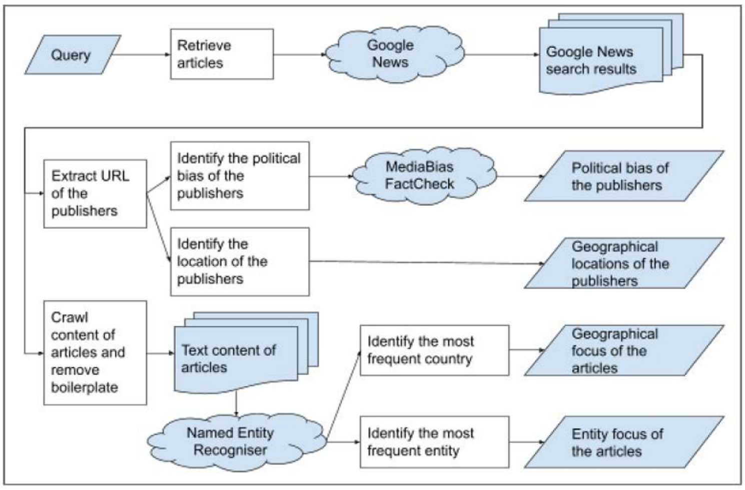

Query
Retrieve articles
Google News
Google News search results
Extract URL of the publishers
Identify the political bias of the publishers
MediaBias FactCheck
Political bias of the publishers
Identify the location of the publishers
Geographical locations of the publishers
Crawl content of articles and remove boilerplate
Text content of articles
Identify the most frequent country
Geographical focus of the articles
Named Entity Recogniser
Identify the most frequent entity
Entity focus of the articles

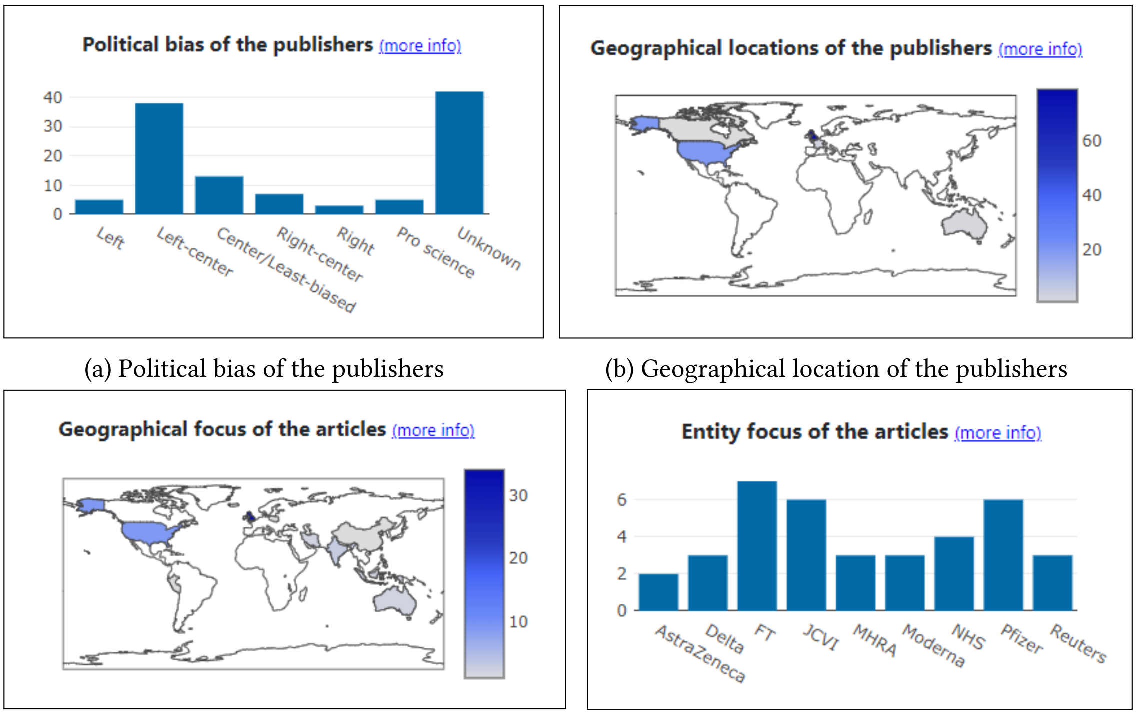
Figure 2: BASE Workflow

Secondly, we crawled the content of the articles and removed the boilerplates. These contents were processed using a named entity recogniser to identify the geographical focus of the article and the entity focus of the article. We describe these processes in more detail below.

Political bias. We utilised an external resource, Media Bias/Fact Check (MBFC) [10], to identify the political bias of the publishers. MBFC is an online source that provides annotations of biases based on the publishers' i) political affiliations, ii) story choices (if they publish from both sides or just one), iii) use of biased wording to sway readers, and iv) rates of factual reporting. By August 2021, MBFC has annotated 3,103 news publishers using five different rating to represent the political bias: "left" represents a liberal view, "left-center", "center/least biased", "right-center" and "right" represents a conservative view. It has further used four categories to represent sites that are considered to be "questionable sources", "pro-science", "satire" or containing "conspiracy-pseudoscience". These ratings were extracted to represent the "political bias of the publishers" in the prototype. For cases where publishers were not included in the MBFC database, the political bias is listed as "unknown". The political bias of all the news articles displayed in the results is aggregated and represented in a bar chart (see Figure 3a) to allow readers to get some insights into the possible bias presented in their search results.

Geographical location of the publishers. We determined the location of the publishers by analysing the suffix of the URL (e.g., "`bbc.co.uk`" is based in the UK, "`abc.net.au`" is based in Australia). When this information was not available, we used '`whois`' command to identify the country where the domain is registered. Similarly, the publisher location was extracted for each article, and was aggregated for all the search results. This information is displayed using a choropleth map (see Figure 3b).

Geographical focus of the articles. We used Scrapy [11], an open-source web-crawling framework, to crawl the content of the articles. Boilerplates were removed using jusText library [12], resulting in the main text content of the articles. We used a named entity recogniser (spaCy [13] trained using the en_core_web_trf model) to identify country names discussed in each



(a) Political bias of the publishers

(b) Geographical location of the publishers

(c) Geographical focus of articles

(d) Entity focus of articles

Figure 3: Bias visualisation features

article. The most frequent country is selected as the geographical focus of the article. Similarly to the locations of the publishers, this information is also aggregated at the results level and is visualised using a choropleth map (see Figure 3c).

Entity focus of the articles. We used spaCy [13] to identify the most frequent entities discussed in the article. If multiple entities had the same frequency, one was chosen randomly as the entity focus of the article. This information was aggregated for all the search results and shown in a bar chart. E.g., Figure 3d shows the most popular entities for the query: "covid vaccine". This includes popular vaccine manufacturers, such as "Pfizer", "Moderna" and "AstraZeneca", and also relevant UK government and health entities such as "Joint Committee on Vaccination and Immunisation" (JCVI), "Medicines and Healthcare products Regulatory Agency" (MHRA) and "National Health Service" (NHS).

2.3. Infrastructure

Due to the amount of processing required, the bias identification task was performed offline. Once completed, the bias information (and focus) was stored in an index, together with each article's information (e.g., URL, title, snippets, etc.). When users submit a query to the BASE system, the articles are retrieved and displayed on the graphical user interface. The interface is developed using PHP, and the visualisations (bar chart and choropleth map) are developed using Plotly Javascript open source graphing library [14].

2.4. Preliminary Evaluation

A preliminary evaluation study involving 21 participants – 47.62% BSc, 33.33% MSc and 4.76% PhD students, and 14.29% non-students; 38.1% males and 61.9% females; ranging from 18 to over 40 years old; from Cyprus (42.9%), Greece (47.6%), France (4.8%) and Italy (4.8%) – suggested that this prototype achieves a moderately positive usability score (64.3 out of 100 using the System Usability Scale) [15]. Some participants mentioned that the system provided too much information that might be too complex for some to use. However, other participants found the system to be easy to use and had the potential to provide more transparency of search results.

A further evaluation study involving 60 MSc students – 55% males, 43.33% females, 1.67% preferred not to say; 92% between 21-25 years old, and the remaining 26 and older; majority (88.33%) from China, and the rest from other Asian countries and Slovakia – suggested that they found the reranking results features to be the most useful (4.08 out of 5). Bias information at the results level were found to be more useful (4.02) than those at the article level (3.75), due to the difficulties to understand the meaning of bias icons for each article (left panel). Participants liked the distribution of biases in the search results (right panel). They also liked the ability to click on the bar chart or maps to easily retrieve results from each aspect. Further feedback from users suggested that users need more clarity, especially how biases were calculated. Others also suggest that the design should be more inclusive, as the "left" and "right" aspect for political ideologies are not necessarily the same nor a familiar concept for users from other countries.

3. Reflections and conclusions

We realise that bias identification is a challenge on its own and may contain its own subjectivities and biases. We reduced this risk by selecting trustworthy resources (MBFC) and focusing on biases that can easily be determined (e.g., locations). MBFC, however, does not have an extensive coverage, especially for non-English news sites. Moreover, the named entity recogniser does not map any cities or towns towards the relevant country counts. It also selects the most frequent entities without taking the query context into account (e.g., that Pfizer and Moderna are relevant entities for "covid vaccine" query, but Reuters is not). More sophisticated methods, therefore, will need to be implemented to accurately identify biases in news search results.

Despite these limitations, the BASE prototype illustrates how biases in search results could be communicated to the users. The prototype incorporates bias visualisation and results-reranking features to inform users of the existing biases and support them in their search tasks. We utilised available resources and NLP tools to identify biases in search results. Our initial evaluation shows that the prototype has potentials for increasing transparency of search results. Future work will investigate ways to improve these features and to reduce the complexity of the system.

Acknowledgments

We would like to thank the reviewers for their constructive feedback on this paper. This project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 810105 (CyCAT).

References

[1] F. Hamborg, K. Donnay, B. Gipp, Automated identification of media bias in news articles: an interdisciplinary literature review, International Journal on Digital Libraries 20 (2019) 391–415. URL: https://doi.org/10.1007/s00799-018-0261-y. doi:10.1007/s00799-018-0261-y.

[2] M. L. Paramita, K. Orphanou, E. Christoforou, J. Otterbacher, F. Hopfgartner, Do you see what I see? Images of the COVID-19 pandemic through the lens of Google, Information Processing & Management 58 (2021) 102654. URL: https://www.sciencedirect.com/science/article/pii/S0306457321001424. doi:10.1016/j.ipm.2021.102654.

[3] A. Novin, E. Meyers, Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 175–184. URL: https://doi.org/10.1145/3020165.3020185. doi:10.1145/3020165.3020185.

[4] R. Epstein, R. E. Robertson, The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections, Proceedings of the National Academy of Sciences of the United States of America 112 (2015) E4512–4521. doi:10.1073/pnas.1419828112.

[5] AllSides | Balanced news via media bias ratings for an unbiased news perspective, 2022. URL: https://www.allsides.com/unbiased-balanced-news.

[6] Ground News, 2022. URL: https://ground.news/.

[7] F. Hamborg, N. Meuschke, B. Gipp, Matrix-Based News Aggregation: Exploring Different News Perspectives, in: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, pp. 1–10. doi:10.1109/JCDL.2017.7991561.

[8] P. Papadakos, G. Konstantakis, bias goggles: Graph-Based Computation of the Bias of Web Domains Through the Eyes of Users, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 790–804. doi:10.1007/978-3-030-45439-5_52.

[9] Zenserp, 2022. URL: https://zenserp.com/.

[10] Media Bias/Fact Check, 2022. URL: https://mediabiasfactcheck.com/.

[11] Scrapy, 2022. URL: https://scrapy.org/.

[12] jusText, 2022. URL: https://pypi.org/project/jusText/.

[13] spaCy, 2022. URL: https://spacy.io/models/en.

[14] Plotly javascript open source graphing library, 2022. URL: https://plotly.com/javascript/.

[15] System usability scale (sus), 2022. URL: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.